# Practical Machine Learning: Course Project

PREDICT 422 – SECTION 56

Winter 2017

Chris Anderson

## Introduction

This report discusses the application of machine learning to improve the cost-effectiveness of direct marketing campaigns for a charitable organization. Given the overall response rate of 10% based on historical records, contacting every individual in the donor database is inefficient and financially irresponsible. Under this strategy, the expected return is actually a loss due to the cost associated with each mailing. In order to optimize the mailing process and improve the expected return, a classification model to identify likely donors will be developed using data from the organization's most recent campaign. Furthermore, an additional predictive model will be developed to predict the expected gift amounts from individuals identified as likely donors. The combination of the two models will allow the organization to plan its next marketing campaign in a manner that increases donations but decreases cost.

The dataset includes over 6,000 records from a recent campaign broken into training, validation, and test sets. In addition to the dependent variables that identify donors and the amount of their donations, it includes 20 independent variables for each observation. These represent various demographic characteristics of an individual, as well as details on his/her previous contributions to the organization. These variables, or transformations of them, may or may not be included in the models. The analysis will be performed using the statistical software R (code included as a separate file).

## Analysis

Data exploration is a critical component in the model building process. It is important to understand the data, handle any missing data or potential outliers, and transform predictor variables if necessary before developing a predictive model. The charity dataset consists of 23 variables representing various characteristics of the potential donors in the organization's records. The variables are defined in Table 1 below.

***Table 1: Variable definitions***

| Variable Definition | | | |
|---|---|---|---|
| **ID** | Identification number | **INCA** | Average family income in individual's neighborhood ($ thousands) |
| **REG1** | Dummy variable representing one of five geographic regions | **PLOW** | Percentage of low income families in individual's neighborhood |
| **REG2** | Dummy variable representing one of five geographic regions | **NPRO** | Lifetime number of promotions received to date |
| **REG3** | Dummy variable representing one of five geographic regions | **TGIF** | Dollar amount of lifetime gifts to date |
| **REG4** | Dummy variable representing one of five geographic regions | **LGIF** | Dollar amount of largest gift to date |
| **HOME** | Dummy variable indicating if individual is a homeowner | **RGIF** | Dollar amount of most recent gift |
| **CHLD** | Number of children | **TDON** | Number of months since last donation |
| **HINC** | Household income rating on scale of 1 to 7 | **TLAG** | Number of months between first and second donations |
| **GENF** | Dummy variable indicating if individual is female | **AGIF** | Average dollar amount of gifts to date |
| **WRAT** | Wealth rating on a scale of 1 to 9 | **DONR** | Dummy variable representing donors (classificiation response variable) |
| **AVHV** | Average home value in individuals's neighborhood ($ thousands) | **DAMT** | Donation amount in dollars (prediction respones variable) |
| **INCM** | Median family income in individual's neighborhood ($ thousands) | *Additional variables may be derived from this list* | |

Each variable will be referenced throughout this report by its name in the table. The variables reg1, reg2, reg3, reg4, home, chld, hinc, genf, and wrat are qualitative in nature. The variables

avhv, incm, inca, plow, npro, tgif, lgif, rgif, tdon, tlag, and agif are quantitative. Donr and damt are the target variables for the predictive models.

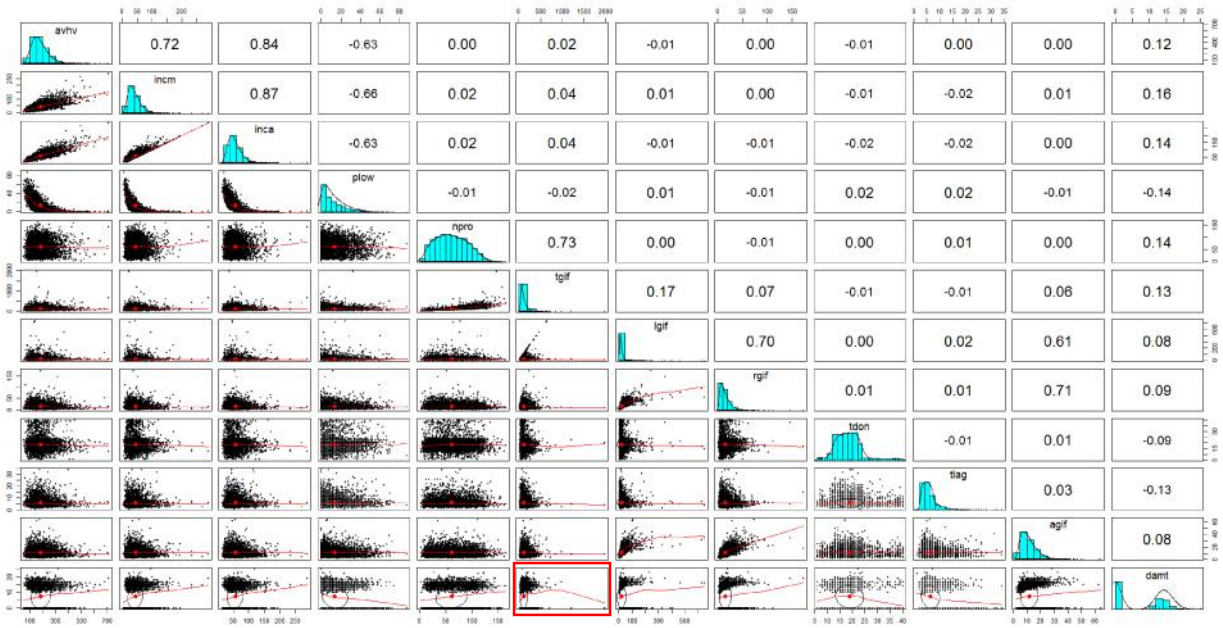
First, it is important to check the completeness of the dataset by examining any missing data. Fortunately, there are no missing observations in any part of the dataset – training, validation, or test. Simple descriptive statistics can also be useful in understanding quantitative variables. Often, it is easy to identify extreme outliers or data entry errors (e.g. negative numbers or extremely large numbers that do not make sense) by looking at these statistics, which are presented below in Table 2.

***Table 2: Descriptive statistics for quantitative variables***

| avhv | incm | inca | plow | npro | tgif | lgif | rgif | tdon | tlag | agif | damt |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Min. : 54.0 | Min. : 3.00 | Min. : 15.00 | Min. : 0.00 | Min. : 2.00 | Min. : 25.0 | Min. : 3.00 | Min. : 1.00 | Min. : 6.00 | Min. : 1.000 | Min. : 1.89 | Min. : 0.00 |
| 1st Qu.:134.0 | 1st Qu.: 27.00 | 1st Qu.: 40.00 | 1st Qu.: 4.00 | 1st Qu.: 37.00 | 1st Qu.: 65.0 | 1st Qu.: 10.00 | 1st Qu.: 7.00 | 1st Qu.:15.00 | 1st Qu.: 4.000 | 1st Qu.: 6.99 | 1st Qu.: 0.00 |
| Median :171.0 | Median : 39.00 | Median : 52.00 | Median :10.00 | Median : 60.00 | Median : 91.0 | Median : 15.00 | Median : 12.00 | Median :18.00 | Median : 5.000 | Median :10.22 | Median :10.00 |
| Mean :185.2 | Mean : 44.29 | Mean : 57.14 | Mean :13.73 | Mean : 61.63 | Mean : 116.7 | Mean : 23.19 | Mean : 15.55 | Mean :18.81 | Mean : 6.302 | Mean :11.60 | Mean : 7.26 |
| 3rd Qu.:219.0 | 3rd Qu.: 55.00 | 3rd Qu.: 68.00 | 3rd Qu.:20.00 | 3rd Qu.: 84.00 | 3rd Qu.: 143.0 | 3rd Qu.: 25.00 | 3rd Qu.: 20.00 | 3rd Qu.:22.00 | 3rd Qu.: 7.000 | 3rd Qu.:14.79 | 3rd Qu.:14.00 |
| Max. :710.0 | Max. :287.00 | Max. :287.00 | Max. :87.00 | Max. :164.00 | Max. :1974.0 | Max. :642.00 | Max. :173.00 | Max. :40.00 | Max. :34.000 | Max. :64.22 | Max. :25.00 |

There are no unexpected negative values identified by the minimum values, and the quantiles, medians, and means reveal nothing out of the ordinary. It appears that the organization has kept excellent records of potential donors. However, there are a few variables for which the maximum value appears to be much larger than the rest of the observations. For example, the mean value for tgif is 91, but the maximum value is 1,974. This outlier may have a significant impact on the accuracy of a predictive model. Another way to view potential outliers or issues with the data is by plotting histograms and scatterplots. These are shown below in Figure 1.

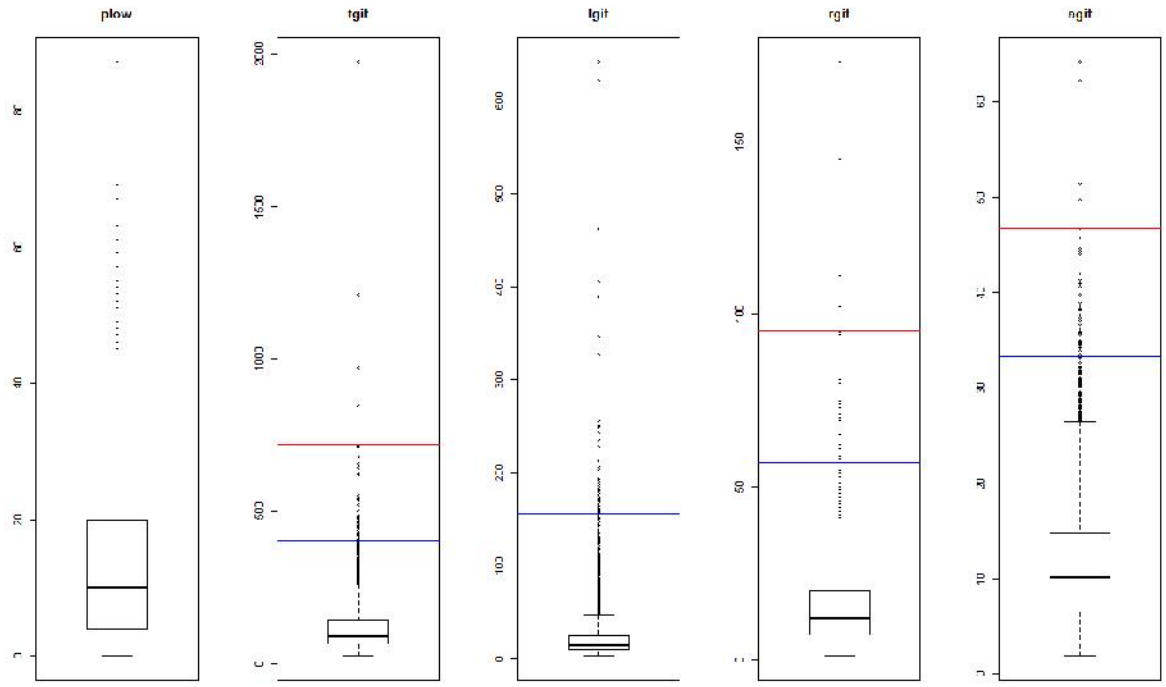
***Figure 1: Correlation matrix with histograms***



The correlation coefficients on the upper panel reveal some unsurprising relationships amongst the predictor variables. For example, there is a strong positive correlation between incm and inca, as well as between lgif and rgif. These relationships make intuitive sense, but it could lead to overfitting combinations of correlated variables are included in the regression model. Furthermore, none of the variables display a particularly strong association with the regression response variable damt. The strongest correlations appear in incm, inca, and npro, though they are still fairly weak. This suggests that a more complex multivariate model may be required in order to explain a significant amount of the variation in damt.

The histograms on the diagonal of the matrix show that many of the predictor variables do not follow a normal distribution, which can cause some statistical models to perform poorly. Many of them are positively skewed, meaning that there are a small number of large values towards the right end of the plot. The plot outlined in red is particularly illustrative of the impact a significant outlier can have. The red line within the plot represents a smoothed trend line. It starts out trending positively upwards and to the right, but it is pulled downwards by a single point in the bottom right corner of the plot. This is the outlier in tgif mentioned on page 2. Removing the outliers may help move the distributions towards normality. Figure 2 below offers a closer look at the outliers in the variables plow, tgif, lgif, rgif, and agif.

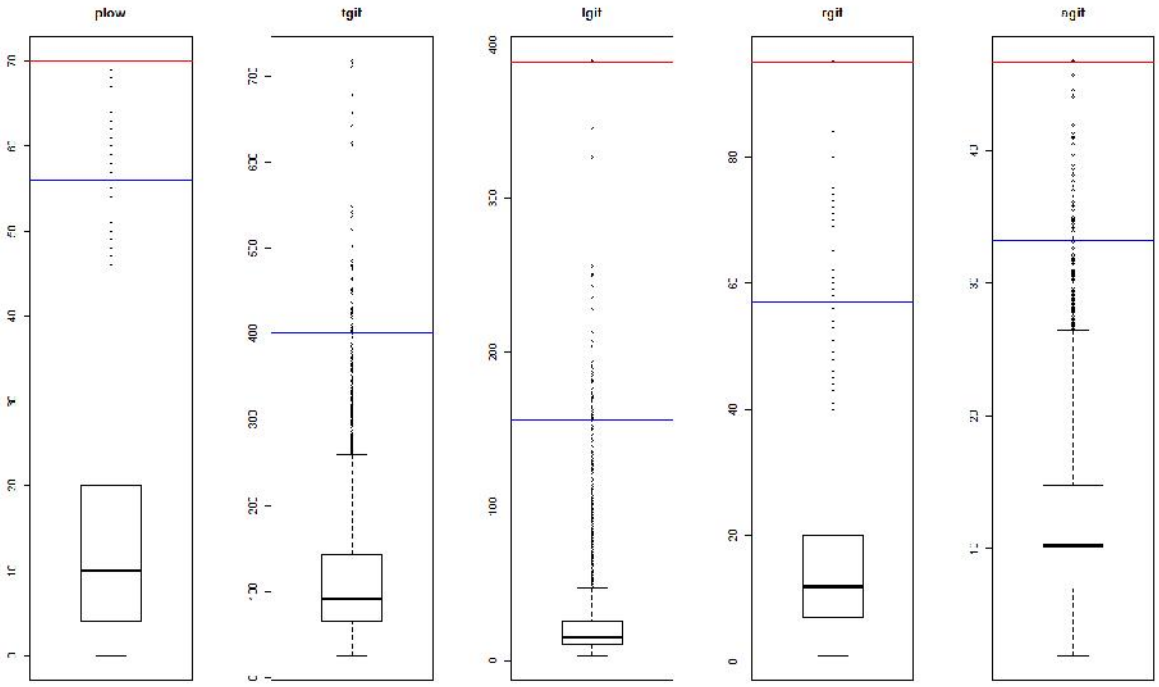
***Figure 2: Boxplots before outlier handling***




There are quite a few observations that fall outside of the interquartile range illustrated by the boxplot. However, labelling all of them as outliers and removing them from the training set may

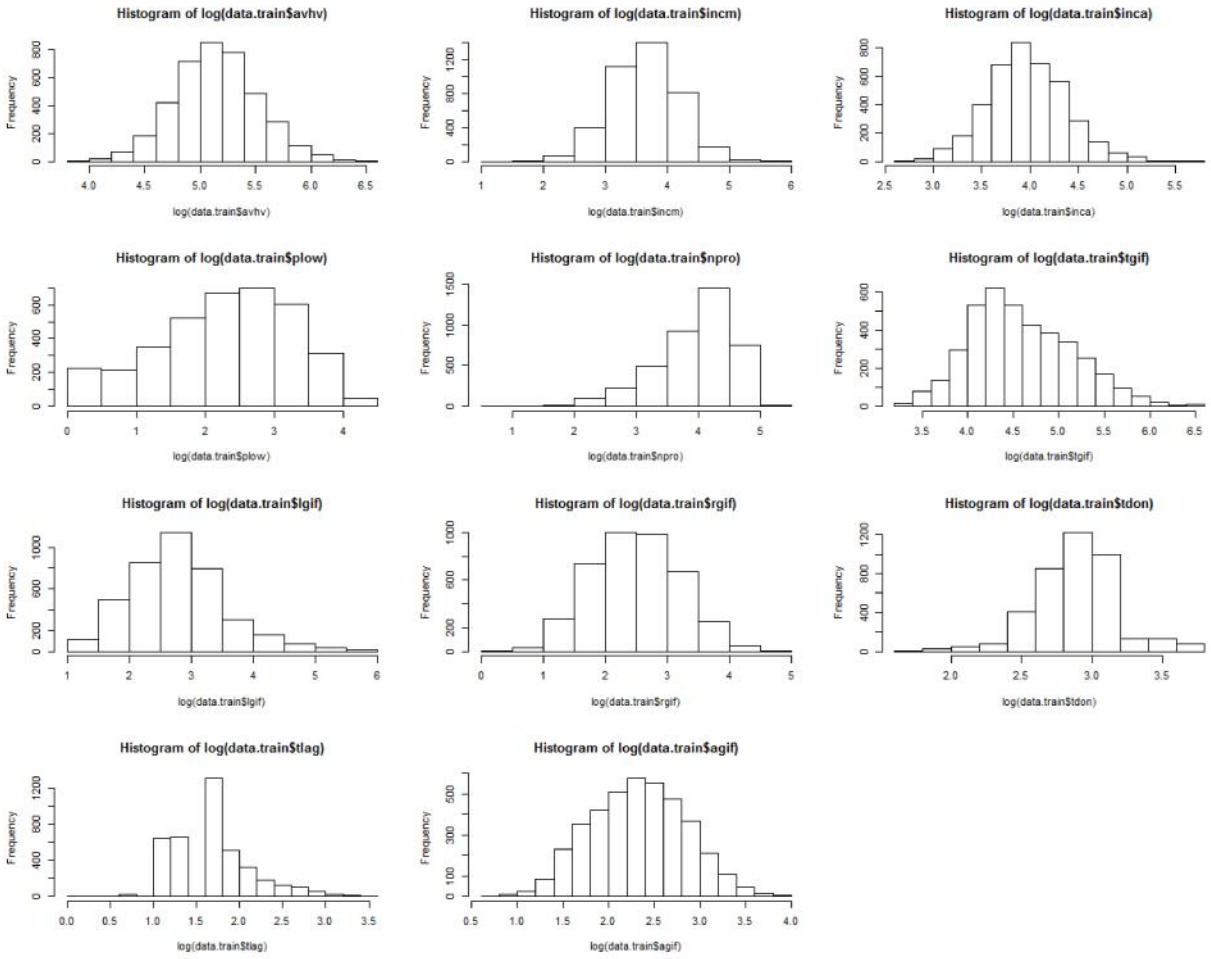
limit the model's predictive accuracy on new data. An alternate method of outlier handling is to trim the most extreme values down to a more reasonable value. In Figure 2, the blue line represents the 99$^{th}$ percentile for each particular variable, and the red line represents the 99.9$^{th}$ percentile. It is a large dataset, so many observations surpass the blue barrier, but only a few of the most extreme values extend past the red line. For tgif, lgif, rgif, and agif, these represent the donors whose gifts have greatly exceeded those of their peers. Figure 3 below shows the same plots after changing the value of any observation that fell beyond the red line to equal the 99.9$^{th}$ percentile.

***Figure 3: Boxplots after outlier handling***




There are still many points that may look like outliers, but further adjustments of this manner may limit the model's ability to predict these high-value donors. Rather, another way to reign in large values is to transform them. Typically, log transformations are useful in these situations when the goal is to convert a positively skewed distribution to a more normal one. Figure 4 below displays log-transformed distributions of the variables that showed positive skewness in Figure 1.

***Figure 4: Log-transformed predictor variables***




The log-transformed variables are mostly normal. Some of the histograms are slightly positively skewed, but mostly there are only minor deviations from normality. However, the plot for the variable npro, in the center panel of the second row, is now negatively skewed. This suggests that the transformation was not necessary, as the distribution is now less normal than it was originally. Also, the plot for the variable plow now looks almost bimodal, and slightly less normal than the original. Therefore, the log transformations will be applied to all variables shown in the above graphic, except npro and plow.

The categorical variables in the dataset require a different type of data exploration. Their relationships with the target variables were explored through cross-reference tables. The following facts were found through this exploration.

- 20.5% of people live in REG1; 55.6% of them are donors
- 33.6% of people live in REG2; 67.4% of them are donors
- 12.3% of people live in REG3; 36.2% of them are donors
- 13.5% of people live in REG4; 34.1% of them are donors

- 88.3% of people are homeowners; 55.3% of them are donors
- 35.0% of people have no children; 86.1% of them are donors
- Household income ratings 3-5 (1-7 scale) are most likely to donate
- Females account for roughly 60% of all donors
- Individuals with a higher wealth rating are more likely to donate

Many of these findings are interesting, but not particularly meaningful. Certain classification methods, such as tree-based methods, will leverage these relationships within the structure of the model without any further action. Two potentially useful statements to explore further are underlined above.

***Table 3: Homeowners with no children***

| donr | home_plus_nokid 0 | home_plus_nokid 1 | Row Total |
|---|---|---|---|
| 0 | 1860 | 129 | 1989 |
| N / Row Total | 0.935 | 0.065 | 0.499 |
| N / Col Total | 0.689 | 0.101 | |
| | 0.467 | 0.032 | |
| 1 | 841 | 1154 | 1995 |
| N / Row Total | 0.422 | 0.578 | 0.501 |
| N / Col Total | 0.311 | 0.899 | |
| | 0.211 | 0.290 | |
| Column Total | 2701 | 1283 | 3984 |
| | 0.678 | 0.322 | |

Building on the fact that 86% of people with no children are donors, a dummy variable was added to classify homeowners with no children. The value highlighted in Table 3 to the left indicates that 90% of these individuals donate to the charity. It is possible that this would have been identified by certain classification methods, but introducing the new variable ensures that this subset of high-likelihood donors is identified and captured by the model.

Another interesting discovery from the classification tables is that there is a much larger proportion of donors in hinc ratings 3 – 5, representing the “middle class” by household income, than the other groups. To take a closer look at this subgroup, another dummy variable was created to identify potential donors that fall into this “middle class”.

***Table 4: Middle class donors***

| donr | midinc 0 | midinc 1 | Row Total |
|---|---|---|---|
| 0 | 869 | 1120 | 1989 |
| N / Row Total | 0.437 | 0.563 | 0.499 |
| N / Col Total | 0.750 | 0.396 | |
| | 0.218 | 0.281 | |
| 1 | 290 | 1705 | 1995 |
| N / Row Total | 0.145 | 0.855 | 0.501 |
| N / Col Total | 0.250 | 0.604 | |
| | 0.073 | 0.428 | |
| Column Total | 1159 | 2825 | 3984 |
| | 0.291 | 0.709 | |

The middle class accounts for over 70% of potential donors. Table 4 to the left shows that over 60% of these middle class individuals are donors. This is significant, considering that only 25% of people outside of the middle class are donors. The fact that the relationship between income and the probability of donating is not linear could be an important consideration during model building.

All of the above transformations and derived variables were applied to the entire dataset, including the validation and test sets. At this point, the dataset is cleaned, explored, and prepared for model building. Next, a classification model is required to identify likely donors.

Classification models to identify likely donors were built using the following methods: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic regression, logistic regression GAM, k-nearest neighbors (KNN), classification trees, pruning, bagging, random forests, boosting, and support vector machines (SVM). They mostly included all possible predictor variables, though other combinations were tried throughout the process. Cross validation was used to select certain parameters where appropriate – namely, to select the cost and gamma values for the support vector machine models. All models were checked against the validation set for predictive accuracy. The results from the best model produced by each method are reported in the Results section.

Predictive models to estimate the donation amount were built using the following methods: least squares regression, ridge regression, lasso regression, principal components regression (PCR), partial least squares regression (PLS), regression trees, bagging, random forests, and boosting. Variable selection was performed both manually and systematically. A linear regression model was fit using predictors identified by the variable importance plot resulting from a random forest regression tree. Best subset selection with k-fold cross-validation was performed to select an optimal subset of predictors. Where appropriate, such as in PCR, PLS, and the various tree methods, models were built using all potential predictors. All models were checked against the validation set for predictive accuracy. The results from the best model produced by each method are reported in the Results section.

**Results**

Ultimately, 14 different types of classification models were developed. None of these methods performed egregiously worse than the others, but some did produce more accurate predictions on the validation set. Figure 5 below shows the relationship between the number of mailings and the expected profit for each model. The GAM logistic regression model produced the same results as the first logistic regression model, so it is not included. Similarly, the pruned classification tree did not produce any sort of improvement over the complete tree, so it is also excluded. Some models were more efficient; they produced a relatively high profit with fewer mailings. These may be preferable in a business scenario where there are limitations on the resources available for distributing the mailings. However, in this case, the expected profit is the driving force behind model selection.

***Figure 5: Mailings vs Profit for classification models***

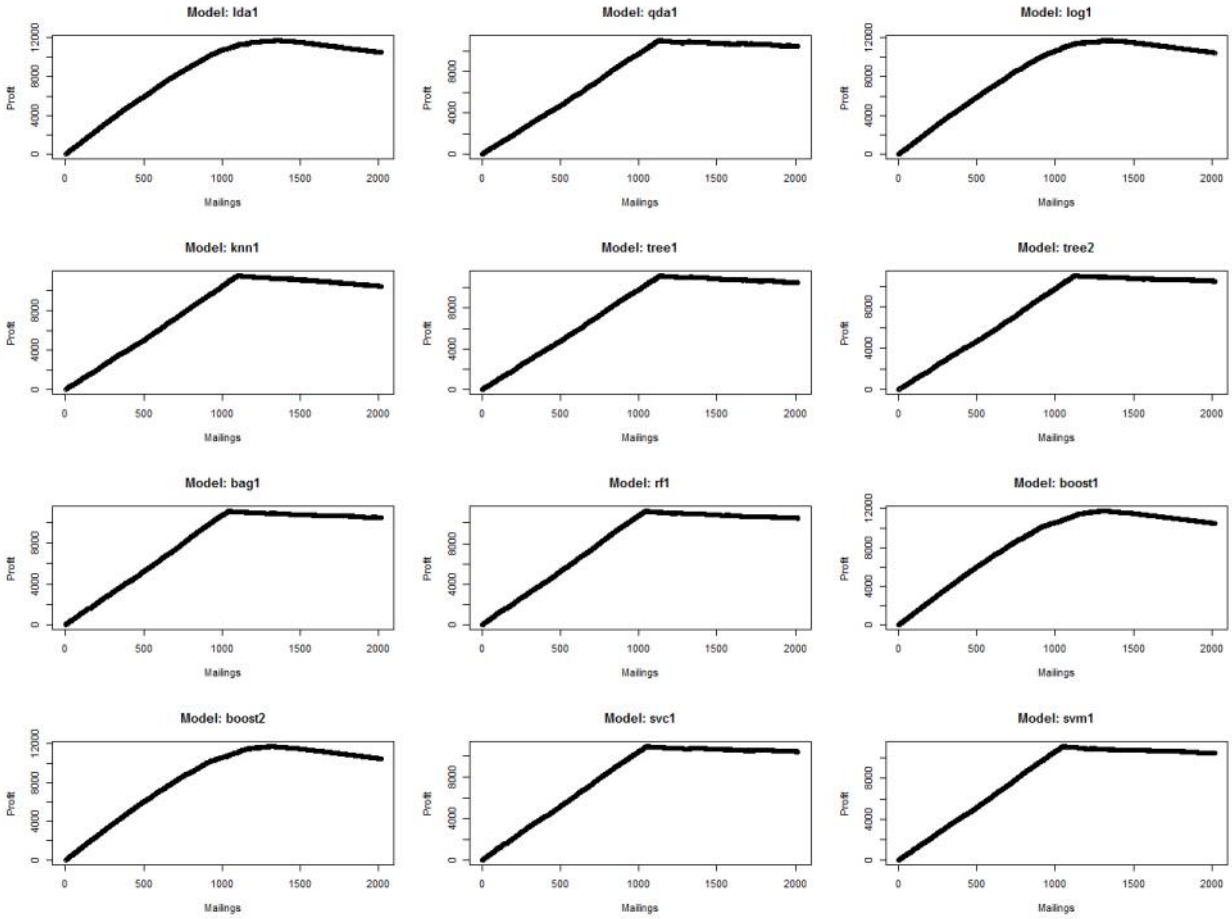



Table 5 below shows the number of mailings that corresponds to the maximum profit for each classification model. The table is sorted by Profit and shows the optimal mailing strategy as determined by each model.

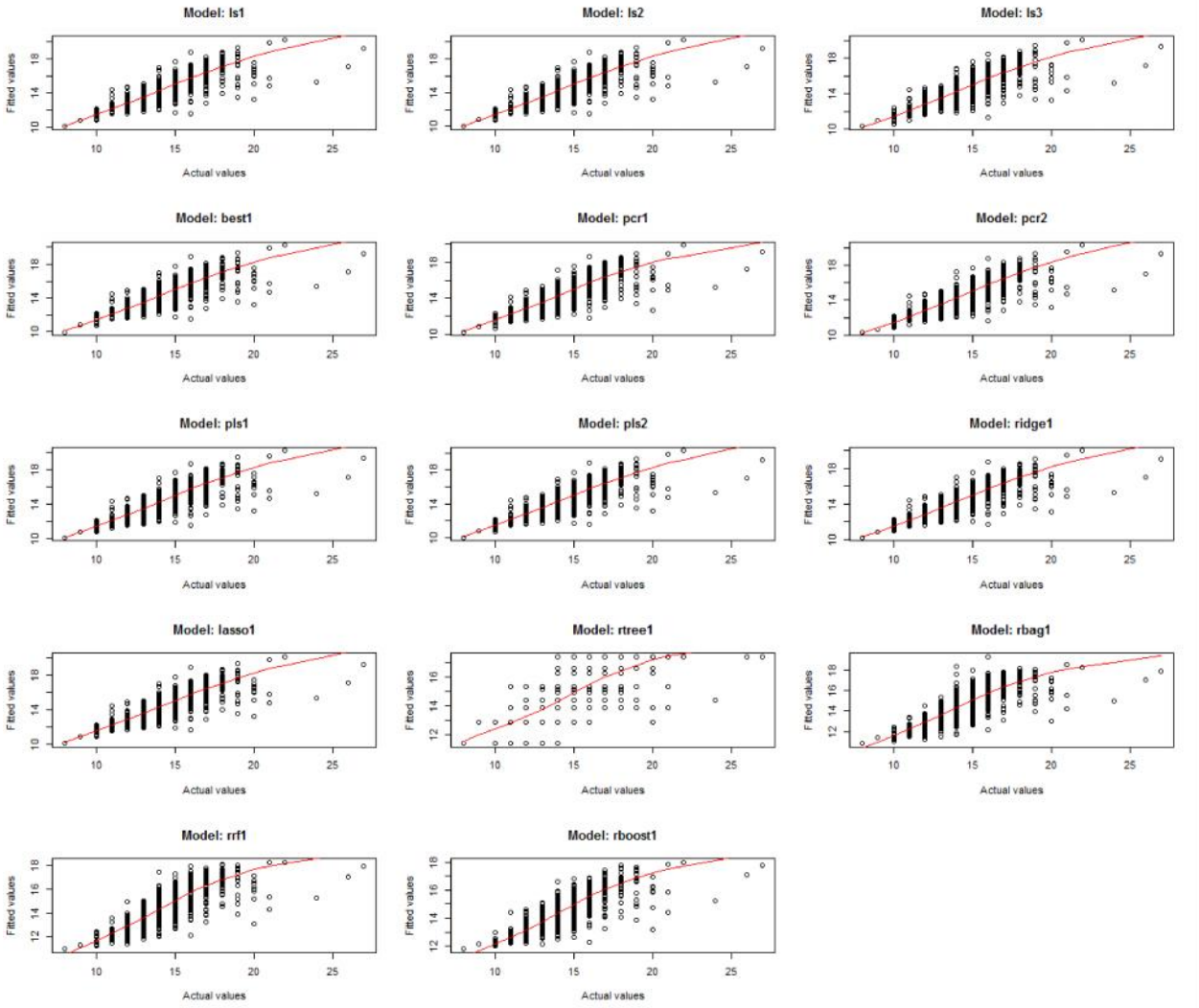
***Table 5: Validation set results***

| Model | Mailings | Profit |
|---|---|---|
| boost1 | 1331 | $11,722.00 |
| boost2 | 1310 | $11,720.50 |
| log1 | 1307 | $11,683.00 |
| gamLR1 | 1307 | $11,683.00 |
| lda1 | 1355 | $11,659.50 |
| knn1 | 1103 | $11,467.50 |
| bag1 | 1055 | $11,186.50 |
| rf1 | 1046 | $11,146.50 |
| tree1 | 1136 | $11,126.00 |
| svm1 | 1050 | $11,124.00 |
| qda1 | 1129 | $11,009.50 |
| tree2 | 1126 | $10,957.50 |
| ptree2 | 1126 | $10,957.50 |
| svc1 | 1054 | $10,956.50 |

The boosted tree classification models performed best on the validation set, followed closely by the logistic regression models. The random forest method produced the most parsimonious model in terms of mailing distribution. This model would lead to the fewest mailings, but would not be expected to produce the most profit. Because it produced the highest expected profit, model “boost1”, a gradient boosted tree using the Bernoulli distribution, is selected as the winning classification model. After converting the number of mailings to adjust for the higher proportion of donors in the training and validation sets, the model determines that mailing to the 356 highest posterior probabilities is the optimal solution.

Furthermore, 9 different methods, along with various variable selection techniques, were utilized to build 14 different regression models to predict the donation amount. To demonstrate the accuracy of each model on the validation set, Figure 6 below displays a plot of the actual values against the fitted values for each of the 14 regression models.

***Figure 6: Actual values vs Fitted values for prediction models***




The red line represents a smoothed trend line, so a strong model would show a straight line from the bottom left to the top right corner of the plot. Most of these models performed fairly well on the validation set based on the above plots. The only model that stands out is model “rtree1”, and it does so in a negative way. It did not perform nearly as well on the validation set as any of the others. Pruning was attempted, as well, but it did not result in an improved model. A more quantitative look at model performance is provided in Table 6 below. The validation set mean prediction error (MPE) and standard error (SE) are reported for each model.

***Table 6: Validation set results***

| Model | MPE | Standard Error |
|---|---|---|
| ls2 | 1.5425 | 0.160072 |
| pls2 | 1.5432 | 0.160070 |
| ls1 | 1.5434 | 0.160123 |
| best1 | 1.5463 | 0.159534 |
| lasso1 | 1.5535 | 0.160397 |
| pls1 | 1.5631 | 0.159341 |
| ridge1 | 1.5709 | 0.162575 |
| pcr2 | 1.5835 | 0.162353 |
| ls3 | 1.6058 | 0.159735 |
| rrf1 | 1.6534 | 0.171253 |
| pcr1 | 1.6657 | 0.161908 |
| rbag1 | 1.7009 | 0.174637 |
| rboost1 | 1.7603 | 0.173015 |
| rtree1 | 2.2700 | 0.191232 |

Not surprisingly, the more traditional regression techniques such as least squares, principal components, ridge, and lasso perform better than the tree-based methods like bagging, random forests, and boosting. These techniques are better suited for qualitative modelling, while trees are more commonly used for classification. It was somewhat interesting to observe that the best models by MPE were not necessarily the best by standard error. This is evidence of the variation in model accuracy metrics and how they can sometimes lead to different solutions. In this case, the best model by standard error is model "pls1", a partial least squares model fit with 5 principal components. Ultimately, the selection criteria identified for this project is MPE, so the best regression model is model "ls2". Interestingly, this model was created by fitting a full least squares model with all predictors, then manually removing those that were flagged as insignificant.

**Conclusions**

The two winnings models – the gradient boosted classification tree and the least squares regression model with manual variable selection – are used to predict the estimated value of the charity's next mailing campaign. This optimal solution consists of 356 mailings, which would produce an expected return of $5,146.80. The average donation is estimated to be $14.46, which aligns with the organization's forecast of $14.50.

Though the dataset was very clean to begin with, a thorough data exploration process helped improve the accuracy of the subsequent models. The use of a validation set allowed for the comparison of a wide range of models, derived using different statistical techniques and different combinations of predictor variables. This process allows the organization to proceed with confidence that the winning model will produce increased donations at a lower cost.